Polishing:
Enhancing Data Quality by Repairing Imperfections

Choh Man Teng

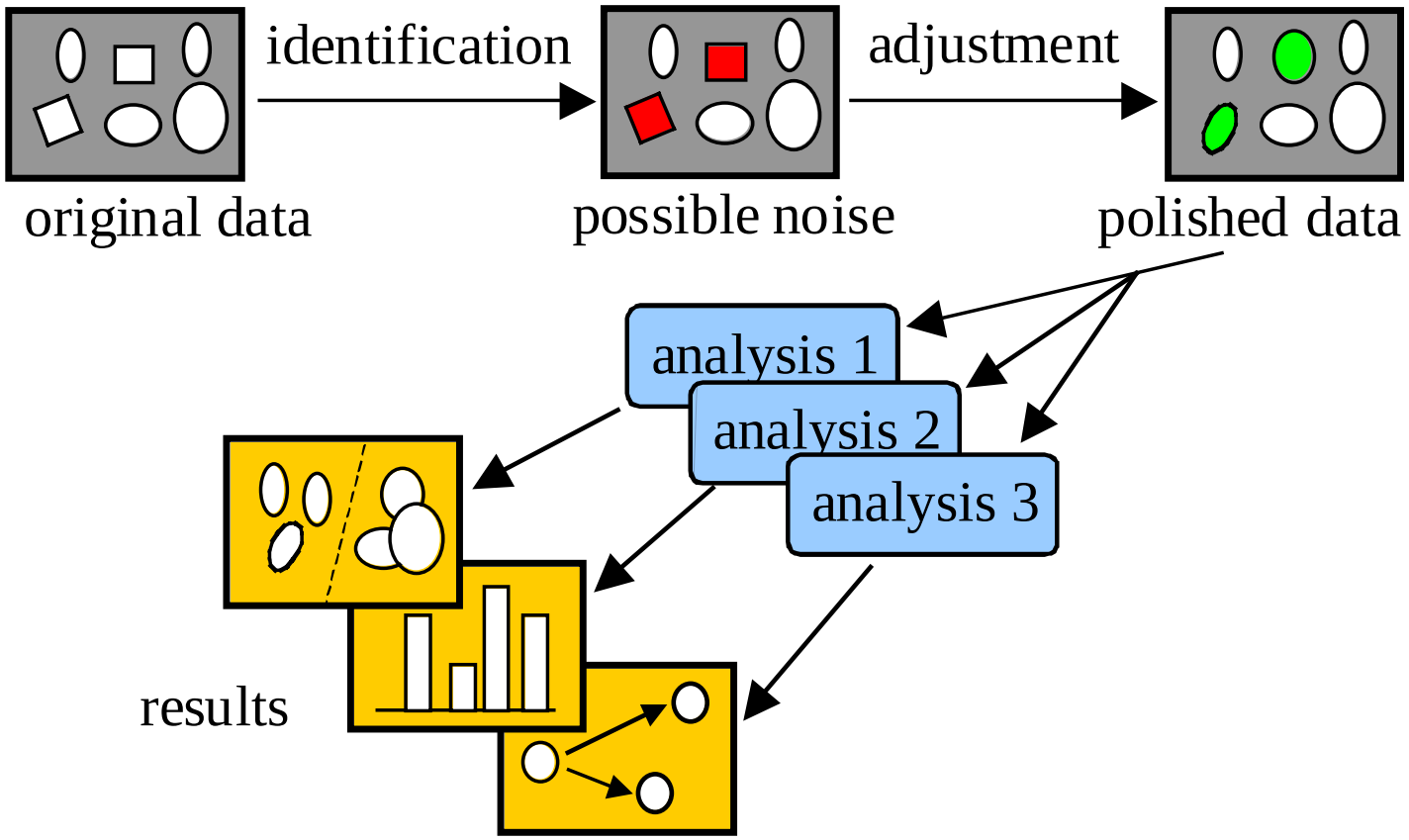
The Problem

- Causes of data imperfections
  - Instrument failures
  - Less than ideal observation conditions
  - Recording and archival constraints
- Imperfect data are less useful than clean data but still contain valuable information
- In many cases perfect data are unobtainable

Objectives

- Identify possible imperfections in the data and correct the problematic portions
  - As opposed to discarding the corrupted portions wholesale
- Advantages
  - Less wastage: Better data, and more of them
  - Higher quality results from higher quality data
  - Recover information otherwise not available

Overview


identification
adjustment
original data
possible noise
polished data
analysis 1
analysis 2
analysis 3
results

Polishing

- Make use of the dependency relationship between different parts of the data
- Prediction
  - Identify corrupted parts and suggest correction
  - Predictions based on models built from original data
  - Candidate repairs: derived from disagreement between predicted and observed/recorded values
- Adjustment
  - Selectively carry out the suggested changes
  - Criteria: improvement in overall fitness/accuracy of the model

MODIS Data

(Moderate Resolution Imaging Spectroradiometer)

- Vegetation indices and landcover products
- Example imperfections
  - Instrument failures: missing entire observation cycles
  - Cloud cover: missing/distorted values

Data Assembly

- MODIS products
  - NDVI/EVI (MOD13A2)

    Normalized Difference Vegetation Index
    Enhanced Vegetation Index
  - Landcover (MOD12Q1)
- One year of data from October 2000, every 16 days
- Sampled uniformly from all land tiles
  - excluding non-vegetation pixels (water, snow/ice, etc.)
- Entire observation cycles missing in June 2001
  - need base value for corresponding variables
  - Realigned the data based on seasons (northern/southern hemispheres)

Simulating Data Corruption

- Adding Gaussian noise to each variable
  - Mean: 0; standard deviation: 0-4000

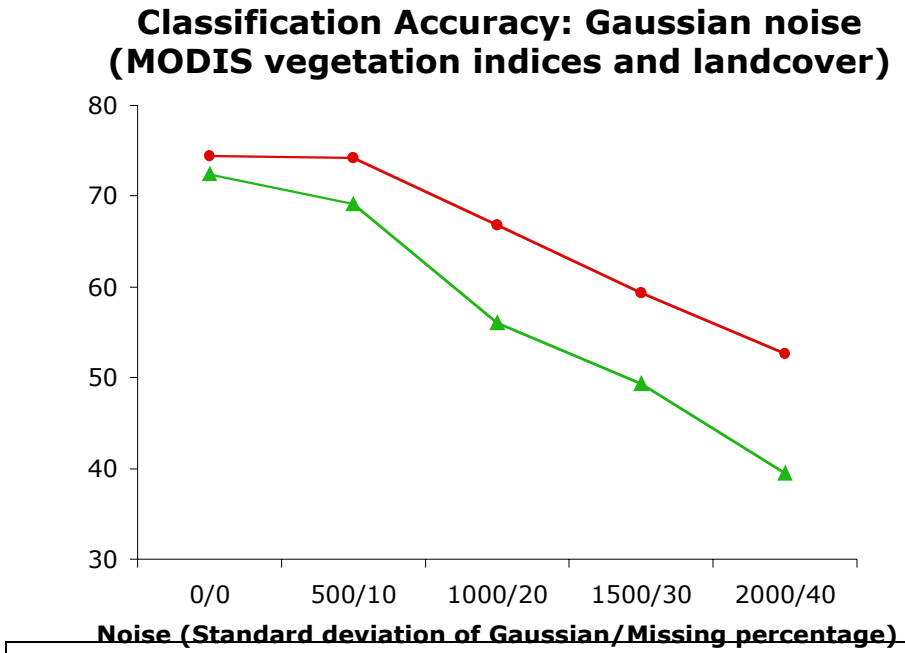
    (range of VI values: -2000 to 10000)
- Knocking out variable values randomly
  - Missing percentage: 0-40%

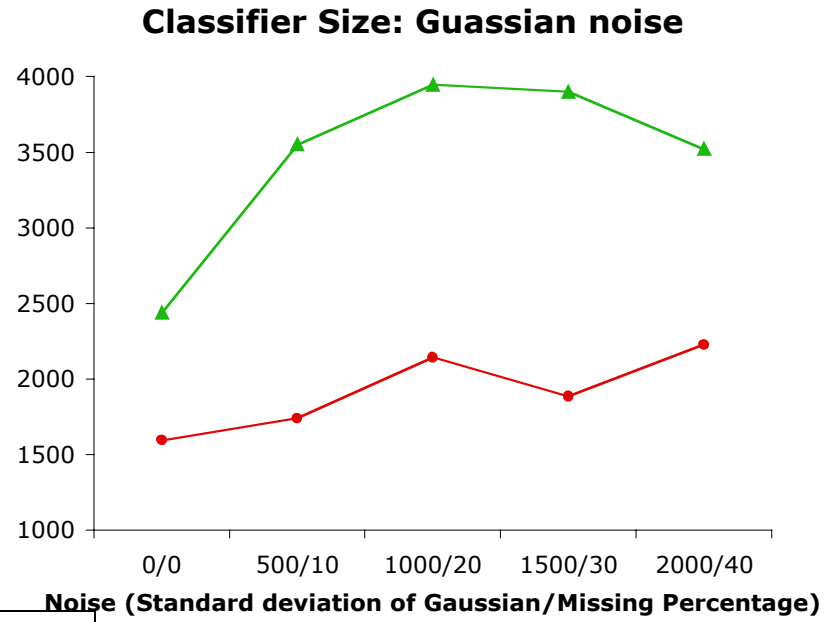
Experiments

- Base classifier: decision trees
- Ten fold cross validation
- Compared unpolished and polished data

Evaluation

- Classification characteristics
  - Categorical accuracy (match/mismatch)
  - Classification confidence
  - Classifier size
- Proximity (x_i: original, y_i: corrupted, z_i: polished)
  - Net reduction in overall corruption
    $\sum_i (|y_i-x_i|-|z_i-x_i|) / \sum_i |y_i-x_i|$
  - Percentage of correct adjustment
    $\sum_{|yi-xi|>|zi-xi|} (|y_i-x_i|-|z_i-x_i|) / \sum_i \big| |y_i-x_i|-|z_i-x_i| \big|$
  - With respect to original data point and nearest neighbor





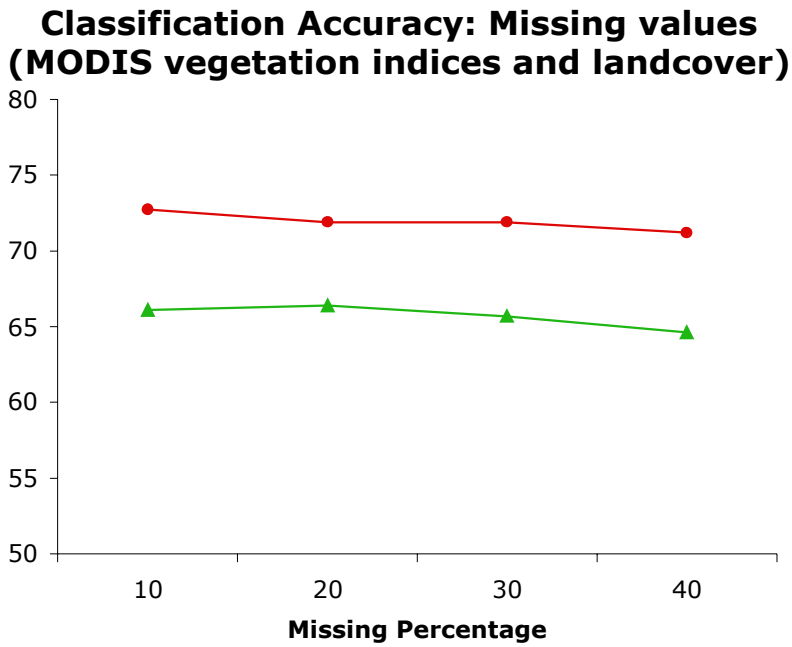
Proximity	Net Reduction (-NN)	(+NN)	Correct Adjustment (-NN)	(+NN)
0/0	---	---	0.0	38.3
500/10	0.7	2.9	53.1	83.8
1000/20	1.9	3.8	61.1	91.5
1500/30	1.6	3.5	61.5	91.6
2000/40	2.3	3.4	71.7	94.4

Some Observations

- A corrupted data point can be adjusted away from the original point but towards another clean data point
- Polishing decreased the overall corruption only a little but most of its adjustments were “correct”, and the adjustments contributed to improved classification performance
- Important to have meaningful characterization of “good” corrections

Missing Values

- Special case of data corruption
  - Corrupted values can be readily located (missing)
- Vegetation indices are related temporally
  - Missing VI's can be filled by interpolation
- Compared interpolation and polishing for filling in missing values


Classification Accuracy: Missing values
(MODIS vegetation indices and landcover)
80
75
70
65
60
55
50
10
20
30
40
Missing Percentage

Some Observations

- Missing values are well tolerated but incorrect values lead to quick deterioration of classification performance
- Perhaps the data set contains highly redundant information (correlated variables)
- Interpolation can flatten useful variations in the pattern
- Effects of: non-random missing patterns; sparser and less redundant data